Amplytica: Bringing microbial ecology to the cloud.

Lee H. Bergstrand, Matt J. McInnes, Kevin O'Neil, Jonathan D. Van Hamme

Department of Biology, Thompson Rivers University

LinkedIn QR

Lee Bergstrand

Background

The **Amplytica Cloud Platform (ACP)** is a software system for building large scale bioinformatics applications on commercial cloud computing infrastructure (Amazon AWS, Google Cloud etc.) with a focus on microbial ecology workloads.

It makes use of emerging open source cloud technologies such as:

- *Docker*
- *Salt-Cloud*
- *RabbitMQ*
- *Binary Large Object (BLOB) stores*
- *Cloud Databases*

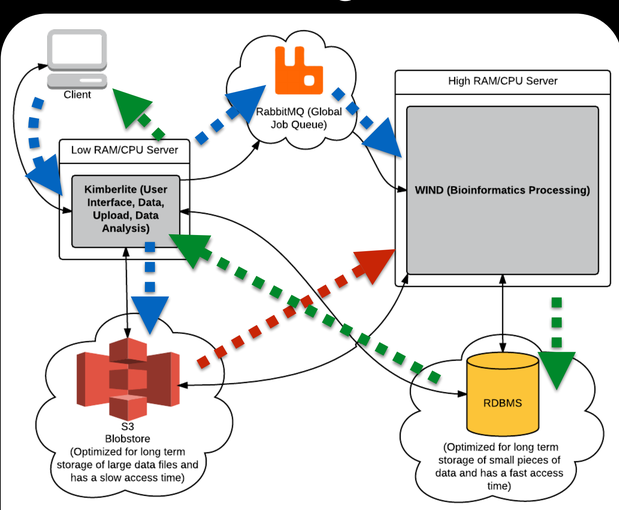
The platform is designed to be distributed across many servers which **do not need to be in a cluster or even on the same public cloud provider.**

Advantages Include:

- *The separation of sequence quality control/analysis from bioinformatics processing allows for different sizes of cloud virtual machines (VMs) to be used for different tasks. This reduces overall costs as it allows VM sizes to be tailored to specific tasks.*
- ***Since the system is distributed, components can be turned on and off on demand allowing end users to pay for only individual bioinformatics processing jobs rather than for longterm servers.***
- *Components are stateless and no sequence data is stored inside them allowing for failure at any time with minimal data loss.*
- *ACP will support multiple sequencer vendors in the near future (it currently supports Ion Torrent).*
- *Can be hosted by Canadian cloud providers or private cloud.*

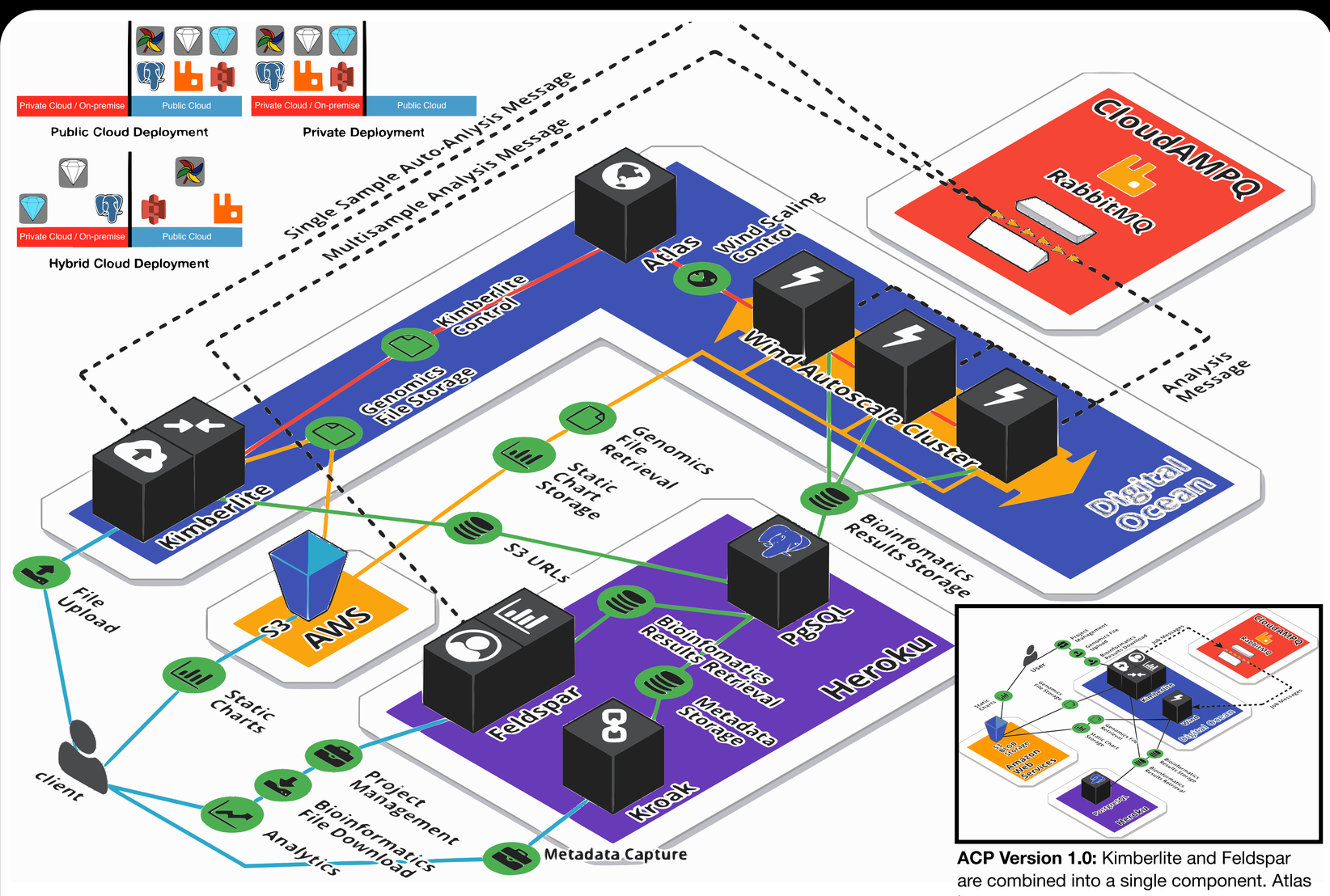
In a microbial ecology context, **ACP processing components wrap the QIIME microbial ecology pipeline allowing it to cluster OTUs on High-RAM cloud virtual machines.** ACP is designed to process, store and organize **thousands** of samples per month.

Design



ACP Version 1.0: The client uploads sequence data to Kimberlite where it is cleaned, compressed and stored in Amazon S3 (**Blue**). Afterwards, a message is sent to Wind indicating that a sample needs processing (**Blue**). Wind responds to the message and pulls the pre-cleaned samples from S3 (**Red**). Wind processes the sample and stores the results in a Relational Database (RDBMS) for later analysis by Kimberlite (**Green**).

Architecture



ACP Version 2.0: Implements metadata capture, auto-scaling and refactors the web interface into its own stand alone service (**in active development**).

ACP Version 1.0: Kimberlite and Feldspar are combined into a single component. Atlas is not present.

Kimberlite:

- Receives sequence file uploads.
- Trims, Error Corrects and Compresses Reads
- Generates QC reports
- Stores compressed sequences files to Amazon S3

Feldspar:

- Serves web interface to users.
- Project Management
- Data Exploration and Analysis
- Sends Analysis Messages to Wind

Wind:

- Asynchronous bioinformatics processing node the responds to analysis messages
- Currently wraps the QIIME pipeline (WIND-QIIME)
- **Can be turned off when processing isn't required.**

Atlas:

- **Turns off Wind VMs when there are no samples to process.**
- **Turns on and autoscales Wind VMs when there are samples to process.**

Kroak:

- Captures project metadata.
- Records metadata datatype for automated analysis
- Excel Compatible

ACP follows a **microservice architecture** where the overall functionality of the software is spread out among individual stand alone services which can be **deployed and updated independently**.

Binary Large Object (BLOB) Store

- Stores sequence data and analysis results.
- Data is placed and retrieved via API.
- ACP currently uses Amazon S3 and is compatible with other BLOB stores (i.e. **Minio**) which use S3's API **(Self-Hostable)**

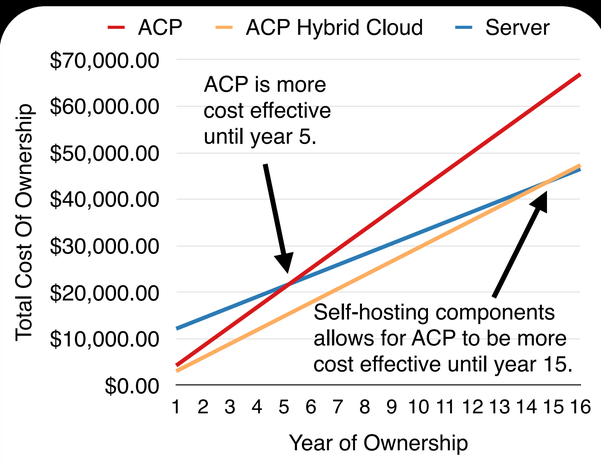
PostgreSQL

- Stores project management data.
- Stores analysis results and integrates data across various services
- ACP currently uses Heroku PostgreSQL, and is compatible with any PostgreSQL database **(Self-Hostable)**

RabbitMQ

- Stores and queues analysis messages between Feldspar and Wind.
- ACP currently uses CloudAMPQ RabbitMQ, and is compatible with any RabbitMQ **(Self-Hostable)**

Results



Total cost of ownership of ACP, ACP with some services self-hosted (Hybrid Cloud) and a $10000 bioinformatics server with undergraduate personnel support (4hrs/Week @ $11/hr) over a 16 year period.

Conclusions

The **ACP** allows for cost effective deployment of bioinformatics software on commercial cloud infrastructure.

- It remains cost effective for at least 5 years, which is around the time where physical servers should be replaced.
- As the cost of cloud computing decreases, ACPs' cost efficiency will continue to increase.
- The platform's design and distributed nature facilitates horizontal scalability and various deployment strategies. Together, these two properties allow for the system to be adapted to fit various budgets, usages and security requirements.

Availability

The ACP is backed by a **for-profit** undergraduate founded microbial bioinformatics **startup** company called **Amplytica Inc.** The platform is currently closed source, however, components will be incrementally open sourced under **Apache License, Version 2.0 (ALv2)** if there is academic and/or commercial partner interest and support. The first component to be open sourced will be **Kimberlite**.

amplytica.io

Amplytica

Support

KAMLOOPS INNOVATION

TRU GENERATOR BCIC

BCIC generously provided an ***Innovator Skills Initiative Grant*** to **Amplytica Inc.** allowing it to employ Matt McInnes as a part time summer student working on the project. Additionally, **Kamloops Innovation** and the **TRU Generator** (Startup Incubators) provided office and co-working space.